Scientific Data Sets: An Important Tool for Student Learning

Sandra Porter, Geospiza, Inc., Seattle, WA

The demands of our increasingly technical society require that today's students develop strong analytical skills and learn how to think in a quantitative manner. One approach to achieving this goal, commonly used in biology labs, has been for students to gather and analyze data sets that they generate on their own through activities such as counting seeds, growing plants, or measuring enzyme activity. This focus on data collection serves an important role in connecting students with data-gathering techniques; however, the labor-intensive and time-consuming nature of these activities prevents students from generating and analyzing large numbers of diverse data sets during the course of a typical 50-minute college class.

One way to provide students with more experience in data analyses is to have students work with pre-compiled data sets, in a convenient form. In this chapter, I will describe some of the benefits to having data sets available, present resources for compiling data sets, discuss advantages of working with authentic data, and discuss features that make data sets useful along with examples of data sets that accompany bioinformatics and biotechnology activities.

Benefits of Data Sets in Biology Education

The rapid development of web-based technologies and the generation of data, through large-scale endeavors such as the human genome project, have pushed biology down the path towards becoming an information-based science. Although this chapter focuses on biology, it seems likely that the availability of data sets and practice in data analysis will be beneficial in other fields as well. With the explosion in information technology, students, in many fields, will need to develop skills for working with data.

In an ideal world, all students would have the opportunity to carry out experiments from start to finish, preparing all of their own samples, collecting all the data, and performing all of the analyses. In the real world, the requirements for lab facilities, technical expertise, and time, along with expensive reagents and costly equipment, make it challenging even for biology students to carry out certain types of experiments and develop adequate collections of data.

Biology courses are a logical place for students to start learning how to work with sets of biological data. However, if biology students are to start working with data types such as protein or DNA sequences, as recommended in the National Research Council's report on biology education (1), they will need to use data obtained by others.

Non-laboratory courses such as math and computer science are faced with even greater difficulties. Many groups, such as NSF (2), the National Institutes of Health (3), the Wellcome trust (4), and others (5,6), have recognized the need for biologists to develop better quantitative skills. At the same time, a growing number of faculty have realized that biology students, in general, are poorly served by existing courses in mathematics. A recent NSF-sponsored workshop at Wake Forest Technical College in North Carolina shed light on some of these reasons. Participants from biotechnology education programs and community college math instructors worked together to develop math problems, many of which used biotechnology data sets (7). Through the workshop, it became clear that the biology instructors were, on the whole, unaccustomed to developing mathematical functions to describe the behavior of the data sets and that the math instructors were unaccustomed to working with

authentic, experimental data. Some of the math instructors were quite surprised by the constraints involved in laboratory work and the methods that biologists use to work around these limitations. Common laboratory techniques, such as the preparation of standard curves, the use of dilutions, or the ubiquitous laboratory practice of using semi-log graphs, were new to the math instructors. Many of the math instructors expressed an interest in using real experimental data sets and introducing biotechnology-related problems into their curriculum.

Unfortunately, most math instructors don't have easy access to the sorts of data that biologists produce. Data sets obtained from diluting and plating cells, protein measurements, cell counts, optical density readings, gel data, and other types of biological assays are not usually available to instructors outside of biology. Thus, the field of biology and biology students are harmed because there are fewer opportunities for students to practice analyzing real data.

Computer science, computational biology, and bioinformatics students are also in need of biological data sets. If students are to learn how to develop algorithms for processing biological data, they need access to multiple types of data sets, including sets with unprocessed data from analytical instruments, such as DNA sequencers and others.

Sources of Data

Although access to some data can still be difficult, this situation appears to be changing. With more data available online and through public databases, educators can now locate many varieties of original data through the web. High-throughput methods of data collection, combined with Internet dissemination, have expanded data accessibility to unprecedented levels. Large databases are now a critical element in biological research and a valuable community resource. Researchers are also increasingly likely to make additional data, supporting peer-reviewed publications, available online. The human genome project has led the biological research community in making data available to others. Not only do genome researchers provide finished data to national databases, they have started depositing data from earlier stages in DNA sequencing (8).

Public databases are the most notable source of research material for compiling data sets. Many of these databases are housed at the National Center for Biotechnology Information (NCBI) (9). All of these resources are experiencing rapid growth, both in volume and usage. Some of the most used databases are GenBank, the primary repository of nucleotide sequence data, with over 2 billion sequences; the non-redundant protein database, containing 1.8 million coding sequences; and the structure database, containing over 25,000 structures of macromolecules. One of the newest databases at the NCBI, the trace archive, exemplifies this growth. In existence since 2001, the trace archive stores over 429,074,410 traces, from several hundreds of different species.

Data sets, accompanying scientific publications, can serve as a valuable resource for education. Not only do students have an opportunity to work with different kinds of authentic data, the activity of deriving their own analyses and comparing their results to a published, peer-reviewed study can be a powerful tool for learning. These types of learning activities, however, have not been widely used. One reason might lie in the contents of research papers. Space limitations impose some constraints on publications, leading most journals to publish only the analyses and not the original, unprocessed data.

Desirable Features of Data Sets

Although large quantities of data can be found in national repositories such as the NCBI and other sources, the data contained within often require additional work to be usable for classroom instruction. Preparing data sets for educational use requires attention to the way that the data set will be used and the sorts of extra information that might be required on the part of the instructor. For example, it would be helpful to many instructors to provide sufficient technical detail so that students and instructors can understand how the data were gathered. Ideally, if data sets are obtained from scientific literature, it's helpful if a full-text version of the original publication is accessible to students online. These give students the opportunity to work with the data first and arrive at their own conclusions. Comparing their results with those from the peer-reviewed publication gives students the opportunity to learn by example and develop their skills through comparison to published authors.

Often, data sets need to be processed so that the material is in a usable form. For example, the PopSet database at NCBI contains sets of sequences, both protein and DNA, derived from population studies. These sequences are ideal

for phylogenetic analysis, but only after several modifications are made. Because the records may contain both protein and nucleic acid sequences, they need to be edited so that the data file only contains one type of sequence. Other information, necessary for the analysis, might not be present in the PopSet record. This information needs to be gleaned from the original publication and provided, either in the sequence file itself or in accompanying materials.

The last problem lies in the programs used to process the data. These programs require that data be organized and sorted in specific ways. One of the most commonly used phylogeny analysis packages, PHYLIP (10), will only allow sequence identifiers that are 10 characters long. So, the identifying information for each sequence in a data set must be re-identified with shorter names, so that sequences can be identified after phylogenetic trees have been produced.

Last, where appropriate, it is helpful to have data sets contain multiple items and answers to questions about the data set. Not only do these data sets allow instructors to assign different examples to each student or group of students, they save the instructor time in not having to solve every problem themselves. Examples of these types of data sets are described below with both structure files and DNA sequences.

Example Data Sets

HIV sequences, sampled over time

A recent data set, developed through this project (DUE 0127599), was derived with the goal of allowing students to prepare and compare phylogenetic trees with an actual timeline (11). The protein sequences in this data set were obtained from a study by Watkins et al. (12), where HIV 1 was grown in cell cultures over a period of 45 weeks and treated with different types of protease inhibitors. The goal of this research was to understand why HIV becomes resistant to multiple protease inhibitors simultaneously. We edited the sequence files in preparation for use with these sequences for use with ClustalX (Figure 1) (13) and PHYLIP (Figure 2) (10). These sequences allow students to prepare different types of phylogenetic trees and compare the results with the time the sample was taken. If a tree presents a likely version of evolutionary events, then HIV sequences obtained earlier in the study should appear in earlier positions on the tree; sequences from later samples should appear later.

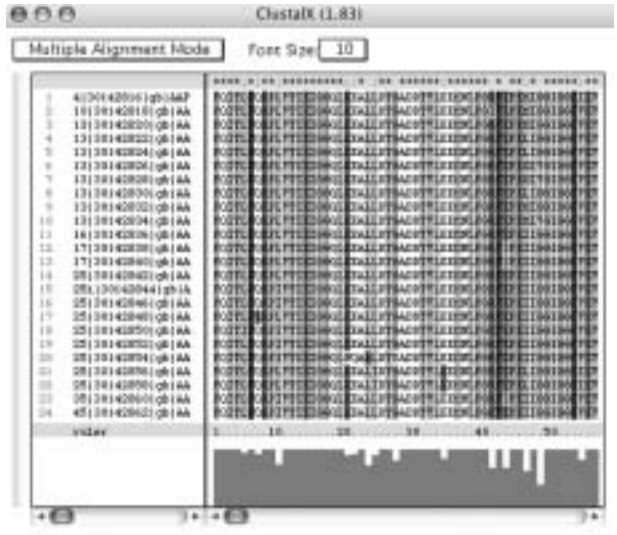

Figure 1. A screen shot from the ClustalX program after performing a multiple alignment with amino acid sequences from an HIV protease. These sequences were obtained from HIV cultured in the presence of a protease inhibitor, indinavir, and sampled at different points of time. Mutations can be identified by the changes in the shading seen in each column.

BLAST for beginners

One of the first data sets developed through this CCLI project (DUE 001853 and DUE 0127599) is a set of sequences that students can use as "unknowns" in querying a nucleotide database with BLASTN (14). These 16 sequences were chosen with the goal of meeting three criteria. They were to represent a diverse set of organisms including humans, bacteria, archea, plants, fungi, and different kinds of viruses. Moreover, each sequence codes for interesting proteins with functions that might be recognized by students, such as a DNA polymerase or alpha amylase. Lastly, none of the sequences contain introns. They were either derived from cDNAs, viruses, or bacteria. This simplifies the analysis because the sequences aren't split by introns. This data set and an accompanying worksheet and tutorial are accessed by approximately 400 people per month, from a diverse set of universities and community colleges. Worksheet answers, for all 16 sequences, are available to instructors upon request.

We have identified one challenge though with this data set. Because the data set represents diverse organisms that have been studied to different extents, some of the questions on the worksheet can be answered for some sequences and not others. For example, questions pertaining to the time of

expression and tissue can only be answered if a specific tissue type is identified in the GenBank record and if the sequence is derived from a multicellular organism. Although the questions were chosen in part to illustrate the fact that different kinds of information are available for different sequences, this can still be disconcerting for instructors new to this field who expect that all questions can be answered.

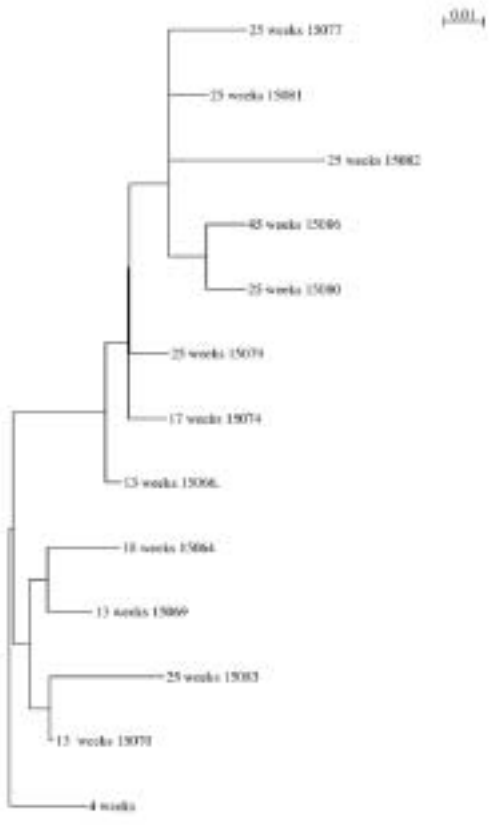

Figure 2. A phylogenetic tree produced from one of the HIV data sets described in the text. The samples that were used for the tree were the amino acid sequences from HIV protease, sampled at different periods of time, while grown in cell culture in the presence of an HIV protease inhibitor, indinavir. The tree was produced with the PHYLIP package of phylogeny software, using the neighbor joining method (Felsenstein, 2004, http://evolution.genetics.washington.edu/phylip/software.html). The tree was produced during the course of testing different phylogeny programs by measuring their ability to reproduce the evolutionary history of an HIV population with the correct chronological order.

DNA groove structures

A data set with 39 different DNA structures (Figure 3), together with a protein or a drug, generated through either nuclear magnetic resonance or X-ray crystallography, was assembled to support a student activity in learning about DNA structure. Each structure shows either a protein or a drug bound to either the major or minor groove of the DNA molecule. In some cases, the structure was annotated to hide a portion of the protein so that only binding to one groove or the other would be shown. This data set is accompanied by a worksheet asking students to determine where binding occurs and an answer sheet for instructors describing each structure and where it binds to DNA. This data set and the accompanying materials are available on a CD from Geospiza, Inc., together with software for viewing the structures (15).

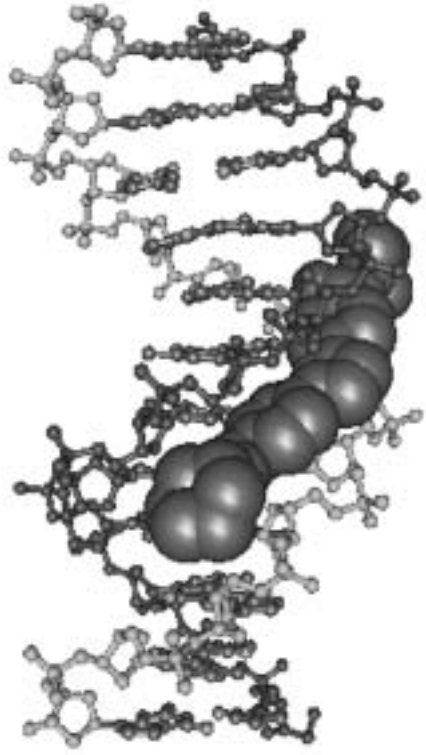

Figure 3. One of the 39 DNA structures that contain a different molecule bound to either the major or minor groove. In this case, the drug molecule, cyclohexyl-bis-furamidine, is binding to the minor groove of a double-stranded piece of DNA. The drug is represented using the space-filling option, and the DNA is drawn with a ball and stick representation. The picture was obtained from Cn3D (16).

Conclusions

Large resource data sets are becoming an increasingly critical component of biomedical and biological research and, as such, will be more frequently produced specifically as community resources. These resources present both opportunities and challenges: challenges because of our inexperience in data processing and opportunities because students have a chance to work with data, develop their analytical skills, and compare their results with experienced researchers.

Experienced researchers and educators can provide students with better opportunities for data analysis by compiling data sets and providing explanatory materials. The availability of these resources will further scientific research by allowing a wider group of instructors to use data analyses as a learning tool.

REFERENCES

1. National Research Council. 2002. *Learning and Understanding: Improving Advanced Study of Mathematics and Science in U.S. High Schools*. Washington, DC: National Academies Press. http://www.nap.edu/catalog/10365.html
2. Clutter, M. 1999. Dear Colleague letter. NSF 99-162.
3. National Institutes of Health Working Group on Biomedical Computing, the Advisory Committee to the Director. 1999. The Biomedical Information Science and Technology Initiative. http://www.nih.gov/about/director/060399.htm
4. Wolfsberg, T. G., K. A. Wetterstrand, M. S. Guyer, F. S. Collins, and A. D. Baxevanis. 2002. A user's guide to the human genome. *Nature Genet* 32 (Suppl.): 1–79.
5. Varmus, H. 2002. Genomic empowerment: the importance of public databases. *Nature Genet* 35 (Suppl.): 3.
6. Baxevanis, A., and F. Collins. 2002. Power to the people. *Nature Genet* 32: 2.
7. A Vision: Technical Math for Emerging Technologies (May 2004). http://www.waketech.edu/~rlkimbal/CRAFTY/may04.html
8. Reaffirmation and Extension of NHGRI Rapid Data Release Policies: Large-Scale Sequencing and Other Community Resource Projects February 2003 NIH-DOE Guidelines for Access to Mapping and Sequencing Data and Material Resources. 1991. www.genome.gov
9. The National Center for Biotechnology Information. http://www.ncbi.nlm.nih.gov
10. Felsenstein, J. 2004. PHYLIP (Phylogeny Inference Package) version 3.6b. Distributed by the author. Department of Genome Sciences. Seattle, WA: University of Washington.
11. Studying Evolution With Bioinformatics, NSF Chautauqua program (2004). http://www.geospiza.com/outreach/chataqua2004/idv_hiv.txt
12. Watkins, T., W. Resch, D. Irlbeck, and R. Swanstrom. 2003. Selection of high-level resistance to human immunodeficiency virus type 1 protease inhibitors. *Antimicrob Agents Chemother* 47: 759–769.
13. Thompson, J. D., T. J. Gibson, F. Plewniak, F. Jeanmougin, and D. G. Higgins. 1997. The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Research* 24: 4876–4882.
14. BLAST for beginners. http://www.geospiza.com/outreach/BLAST.
15. Porter, S. *Exploring DNA, RNA, and Protein Structures: An Introduction to the Molecules of Life*. Seattle, WA: Geospiza, Inc. In press
16. Hogue, C. W. V. 1997. Cn3D: a new generation of three-dimensional molecular structure viewer. *Trends Biochem Sci* 22: 314–316.